Unit 12
End of Module (Numerical Analysis)
Assessment 2
For
M. Sc. In Artificial Intelligence
As required by
University of Essex

Individual Reflection

5th September 2022

The module, Numerical Analysis, helped build and realise the basics, importance, and implementation to create informed decisions with better interpretation.

The following vision stated as my “individual reflection” about the module is outlined according to Rolfe et al.’s (2001) reflective model.

What?

The journey began with learning about Statistics and R. Statistics as the building block for analysis and R as a tool with analytical libraries that can work on large datasets running statistical calculations. During my initial studies, I planned to set up RStudio and learn about R. To make it more interesting; I started building re-usable scripts by testing them on various datasets – from the modules, from the internet like Kaggle, etc. and adding statistical functions & tests that are required for analysis.

I ensured I covered most modules earlier due to upcoming office workloads and vacations, so I did not miss deadlines. This helped when I appeared for the mathematics test. The online testing feature was good, with suggestions for improving the test portal for better HTML support during copy-paste. It becomes difficult for statistics and long ranges of numbers to be sorted if the data cannot be copied.

The seminars were excellent, with lots of collaboration and tutorial involvement. The team was energised and discussed openly. Problems were noted and then resolved with good explanations with real-life examples.

So what?

With a perfect discussion board and tutor & team interactions, we took the initiative in understanding the problems.

To enhance my knowledge of R and the strength of statistics, I looked into datasets to build easy summarised views for quick descriptive analysis using R libraries like –

> dplyr (https://dplyr.tidyverse.org/) as a grammar where a set of verbs can be used to manipulate datasets and frames to overcome complex data manipulation challenges
>
> tplyr (https://atorus-research.github.io/Tplyr/) as a grammar to simplify the datasets and format them for better readability and summary. It is best designed for “clinical safety summaries”.

Used R functions like *as_tibble*, *na.tools*, *read_* (various read libraries),* etc. for formatting data, removing null/NA values, or creating translated columns for analysis. Furthermore made us of hypothesis and statistical analysis tests using *chisq.test*, *t.test*, etc.

Figure 1: Example of removing NA using na.tools

```
[Workspace loaded from ~/.RData]

> library(haven)
>
> # read the data into a data table, X2011HSE
> X2011HSE <- read_sav("/Users/gini/Google Drive/My Drive/01-Self/UoE/RProjects/datafiles/HSE2011.sav")
>
> # using package na.tools replace all NA with 0
> library(na.tools)
na.tools-0.3.1 (2018-06-25) - Copyright © 2018 Decision Patterns
> X2011HSE[is.na(X2011HSE)]=0
> print(X2011HSE)
# A tibble: 10,617 × 58
   hserial   pserial HHSize tenureb      Sex    Age MonthAge   WeekAge PersNo topqual3   HRPID econact  nssec8  Orig
     <dbl>     <dbl>  <dbl> <dbl+l> <dbl+l> <dbl>    <dbl> <dbl+lbl>  <dbl> <dbl+lb> <dbl+l> <dbl+l> <dbl+l> <dbl+
 1 1001011 100101101      1 1 [Own… 2 [Fem…     75       12 997 [Ove…      1 6 [Fore… 1 [HRP] 3 [Ret… 6 [Sem… 1 [WH
 2 1001031 100103101      3 1 [Own… 2 [Fem…     47       12 997 [Ove…      1 4 [NVQ2… 1 [HRP] 1 [In … 1 [Hig… 1 [WH
 3 1001041 100104101      2 1 [Own… 1 [Mal…     77       12 997 [Ove…      1 1 [NVQ4… 1 [HRP] 3 [Ret… 1 [Hig… 1 [WH
 4 1001041 100104102      2 1 [Own… 2 [Fem…     66       12 997 [Ove…      2 1 [NVQ4… 2 [Not… 3 [Ret… 2 [Low… 1 [WH
 5 1001051 100105101      1 1 [Own… 1 [Mal…     44       12 997 [Ove…      1 3 [NVQ3… 1 [HRP] 1 [In … 2 [Low… 1 [WH
 6 1001061 100106101      2 1 [Own… 1 [Mal…     66       12 997 [Ove…      1 1 [NVQ4… 1 [HRP] 1 [In … 1 [Hig… 9 [In
 7 1001071 100107101      1 1 [Own… 1 [Mal…     84       12 997 [Ove…      1 7 [No q… 1 [HRP] 3 [Ret… 4 [Sma… 1 [WH
 8 1001101 100110101      3 1 [Own… 2 [Fem…     63       12 997 [Ove…      1 7 [No q… 2 [Not… 3 [Ret… 3 [Int… 1 [WH
 9 1001101 100110102      3 1 [Own… 1 [Mal…     62       12 997 [Ove…      2 4 [NVQ2… 1 [HRP] 3 [Ret… 3 [Int… 1 [WH
10 1001111 100111101      2 1 [Own… 1 [Mal…     74       12 997 [Ove…      1 2 [High… 1 [HRP] 1 [In … 3 [Int… 1 [WH
# … with 10,607 more rows, and 44 more variables: totinc <dbl+lbl>, eqvinc <dbl+lbl>, NurOutc <dbl+lbl>,
#   relto01 <dbl+lbl>, relto02 <dbl+lbl>, relto03 <dbl+lbl>, relto04 <dbl+lbl>, relto05 <dbl+lbl>,
#   relto06 <dbl+lbl>, relto07 <dbl+lbl>, relto08 <dbl+lbl>, relto09 <dbl+lbl>, Relto10 <dbl+lbl>,
#   Relto11 <dbl+lbl>, Relto12 <dbl+lbl>, ReltoHRP <dbl+lbl>, marstatc <dbl+lbl>, SHA <chr>, gor1 <dbl+lbl>,
#   wt_int <dbl>, wt_nurse <dbl>, SayWgt <dbl+lbl>, SayDiet <dbl+lbl>, htval <dbl+lbl>, wtval <dbl+lbl>,
#   bmival <dbl+lbl>, whval <dbl+lbl>, omdiaval <dbl+lbl>, omsysval <dbl+lbl>, dnnow <dbl+lbl>, totalwu <dbl+lbl>
```

Figure 2: Example of removing NA using dplyr package

```
> library(dplyr)

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union

>
> # convert to local data frame for easy processing
> alcohol_usages <- as_tibble(X2011HSE)
>
> # select specific columns (sex, age, top qualification, total units, marital status, house size, bmi and region)
> usages <- select(alcohol_usages, Sex, Age, topqual3, totalwu, marstatc, HHSize, bmival, gor1, htval, wtval)
> print(usages)
# A tibble: 10,617 × 10
          Sex   Age                     topqual3   totalwu     marstatc HHSize    bmival           gor1 htval wtv
    <dbl+lbl> <dbl>                    <dbl+lbl> <dbl+lbl>    <dbl+lbl>  <dbl> <dbl+lbl>      <dbl+lbl> <dbl> <db
 1 2 [Female]    75 6 [Foreign/other]                0.058 5 [Divorced]       1      25.3 6 [East of En…  162.  66
 2 2 [Female]    47 4 [NVQ2/GCE O Level equiv]        4.99  5 [Divorced]       3       0   6 [East of En…    0    0
 3 1 [Male]      77 1 [NVQ4/NVQ5/Degree or equiv]    49.0   2 [Married]        2      25.6 6 [East of En…  170.  74
 4 2 [Female]    66 1 [NVQ4/NVQ5/Degree or equiv]     0     2 [Married]        2       0   6 [East of En…    0    0
 5 1 [Male]      44 3 [NVQ3/GCE A Level equiv]       30.2   1 [Single]         1       0   6 [East of En…  168.   0
 6 1 [Male]      66 1 [NVQ4/NVQ5/Degree or equiv]    13.6   2 [Married]        2       0   6 [East of En…    0    0
 7 1 [Male]      84 7 [No qualification]             24.6   6 [Widowed]        1       0   6 [East of En…    0    0
 8 2 [Female]    63 7 [No qualification]              0     2 [Married]        3       0   6 [East of En…    0    0
 9 1 [Male]      62 4 [NVQ2/GCE O Level equiv]        4.62  2 [Married]        3       0   6 [East of En…    0    0
10 1 [Male]      74 2 [Higher ed below degree]       47.8   2 [Married]        2       0   6 [East of En…    0    0
# … with 10,607 more rows
>
> # total sample
> total_sample <- nrow(usages)
> print(total_sample)
```

Figure 3: Mathematical calculations

```
>
> # 2. total number of people who drink alcohol
> total_drink_alcohol <- nrow(drinks_alcohol)
> print(total_drink_alcohol)
[1] 7023
>
> # 3. percent who drink alcohol
> percent_drink_alcohol <- (total_drink_alcohol / total_sample) * 100
> print(percent_drink_alcohol)
[1] 66.14863
>
> # 4. total number of males (1) who drink alcohol
> total_males_drink_alcohol <- nrow(filter(select(drinks_alcohol,totalwu, Sex),Sex==1))
> print(total_males_drink_alcohol)
[1] 3247
>
> # 5. percent of males who drink alcohol
> percent_males_drink_alcohol <- (total_males_drink_alcohol / total_sample) * 100
> print(percent_males_drink_alcohol)
[1] 30.58303
>
> # 6. total number of females (2) who drink alcohol
> total_females_drink_alcohol <- nrow(filter(select(drinks_alcohol,totalwu, Sex),Sex==2))
> print(total_females_drink_alcohol)
[1] 3776
>
> # 7. percent of females who drink alcohol
> percent_females_drink_alcohol <- (total_females_drink_alcohol / total_sample) * 100
> print(percent_females_drink_alcohol)
[1] 35.5656
>
```

Used datasets from imagenet, CIFAR-10, KINETIC-700, etc. to explore the data analysis and graph plotting advantages with R using libraries – ggplot2, sunburstR, etc. An example of sunburstR was trying out the visualisation of football events.

Studied and explored further to understand statistics with the basis of “applied tasks” with R to build the descriptive statistical table so that other inferential and other statistical functions can be applied (Pavlenko, Liliia V., et al. "Application of R Programming Language in Learning Statistics." *Proc. 1st Symp. Adv. Educ. Technol*. Vol. 2. 2022.).

Figure 4: Mutating columns and chi-sq tests

```
> # chi test for sex vs drinking status
> df <- usages %>%
+   mutate(drinks=ifelse(usages$totalwu>0,1,2))
> df
# A tibble: 10,617 × 11
         Sex   Age                     topqual3   totalwu     marstatc HHSize    bmival    gor1 htval wtval drir
    <dbl+lbl> <dbl>                  <dbl+lbl> <dbl+lbl>    <dbl+lbl>  <dbl> <dbl+lbl> <dbl+l> <dbl> <dbl>  <dt
 1 2 [Female]    75 6 [Foreign/other]               0.058 5 [Divorced]       1      25.3 6 [Eas…   162.  66.3
 2 2 [Female]    47 4 [NVQ2/GCE O Level equiv]      4.99  5 [Divorced]       3       0   6 [Eas…    0    0
 3 1 [Male]      77 1 [NVQ4/NVQ5/Degree or equiv]  49.0   2 [Married]        2      25.6 6 [Eas…   170.  74.2
 4 2 [Female]    66 1 [NVQ4/NVQ5/Degree or equiv]   0     2 [Married]        2       0   6 [Eas…    0    0
 5 1 [Male]      44 3 [NVQ3/GCE A Level equiv]     30.2   1 [Single]         1       0   6 [Eas…   168.   0
 6 1 [Male]      66 1 [NVQ4/NVQ5/Degree or equiv]  13.6   2 [Married]        2       0   6 [Eas…    0    0
 7 1 [Male]      84 7 [No qualification]           24.6   6 [Widowed]        1       0   6 [Eas…    0    0
 8 2 [Female]    63 7 [No qualification]            0     2 [Married]        3       0   6 [Eas…    0    0
 9 1 [Male]      62 4 [NVQ2/GCE O Level equiv]      4.62  2 [Married]        3       0   6 [Eas…    0    0
10 1 [Male]      74 2 [Higher ed below degree]     47.8   2 [Married]        2       0   6 [Eas…    0    0
# … with 10,607 more rows
> chisq.test(df$drinks,df$Sex, correct=FALSE)

        Pearson's Chi-squared test

data:  df$drinks and df$Sex
X-squared = 2.3797, df = 1, p-value = 0.1229

>
> # chi test for regions vs drinking status
> chisq.test(drinks_alcohol$gor1,drinks_alcohol$totalwu, correct=FALSE)

        Pearson's Chi-squared test

data:  drinks_alcohol$gor1 and drinks_alcohol$totalwu
X-squared = 20967, df = 20520, p-value = 0.01415
```

While analysing data, I came across empirical data (data gathered based on experience). To explore further then started reading articles about the data and their complexities in researching professional development and learning (Nokelainen, Petri, Tahani Z. Aldahdouh, and Alaa A. Aldahdouh. "Bayesian statistics in the research field of professional learning and development." *Methods for Researching Professional Learning and Development*. Springer, Cham, 2022. 213-241).

Now what?

This module has taught me that as I start exploring real and sample datasets by applying the statistical interpretations between variables (changing, continuous etc.), I will be able to visualise the data and build better recommendations. I have learnt to plan and read materials not only what is provided in the module but also by doing a wide reading.

The next step for me is to explore empirical datasets more and then build reusable libraries in R & python for applications.